

On Analytical Methods and Inferences for 2×2 Contingency Table Data from Medical Studies

RICHARD M. ENGEMAN

*Denver Wildlife Research Center, USDA/APHIS, Building 16, DFC, P.O. Box 25266,
Denver, Colorado 80225-0266*

AND

GEORGE D. SWANSON

*Department of Anesthesiology, University of Colorado School of Medicine, Box B110,
4200 East 9th Avenue, Denver, Colorado 80262*

Received September 5, 1990

Analysis of 2×2 contingency tables is not as trivial as it appears. The choice of the statistical test can affect the inferences resulting from data analysis, especially at small sample sizes. Canned statistical programs do not necessarily lead to an appropriate test. These points are demonstrated using examples from the literature. © 1991 Academic Press, Inc.

INTRODUCTION

One of the most seemingly elementary data analyses is for 2×2 contingency table data. These data sets are frequently encountered in medical studies and most investigators feel competent to analyze the data without consulting a statistician. The well-known analytical procedures are contained in the computer outputs for most statistics packages. The best known statistic is probably the Pearson chi-square statistic, which is usually the first statistic given on computer printouts for contingency table analyses. The Yates (*I*) correction for continuity is often used, even though it has been shown to be a very conservative method, particularly for small sample sizes (2-4). The continuity corrected statistic commonly appears with the Pearson chi-square statistic in program outputs.

Special attention should be paid to the small sample size data situations that could cause an investigator the most problems for conducting an appropriate analysis. Applied statistics texts frequently teach that for smaller sample sizes if certain criteria are not met, then Fisher's "exact" test should be applied. For

TABLE 1
DATA FOR SYMPTOM CATEGORIES FROM THE TWO BROILING METHODS AND THE *p* VALUES
FROM THE THREE ONE-TAILED TESTS

Symptom category P = presence A = absence		Mesquite broiling	Gas-flame broiling	Fisher's exact	Pearson chi-square	McDonald <i>et al.</i>
Any respiratory irritation	P	12	8	.011	.005	.005
	A	1	9			
Upper respiratory irritation	P	10	7	.055	.025	.049
	A	3	10			
Lower respiratory irritation	P	11	8	.040	.017	.022
	A	2	9			
Both upper and lower respiratory irritation	P	9	7	.123	.064	.087
	A	4	10			

example, Dixon and Massey (5) recommend using the chi-square only if all expected cell frequencies are greater than or equal to 2, whereas Snedecor and Cochran (6) say to use Fisher's "exact" test if the total sample size is less than 20 or if the total sample size is between 20 and 40 and the smallest expected cell frequency is less than 5. Fisher's test has also been shown to be very conservative (e.g., (2, 4)) and it requires the assumption that all four marginals are fixed. This may not be realistic, but many textbooks recommend that the test should still be used when sample sizes are not appropriate for the Pearson chi-square. Similarly, when the criteria for application of Pearson's chi-square are not met, many computer program outputs will flag those results and recommend the use of Fisher's test, which also is printed. We use two examples from the literature to illustrate some of the analytical methods and inferential problems associated with 2×2 tables.

EXAMPLE 1

The data in Table 1 originates from an article by Johns *et al.* (7) on assessing whether there are negative respiratory effects to mesquite broiler cooks versus gas-flame broiler cooks. The data presented here are corrected data. The data tables in the Johns article originally contained errors which are detected by comparing the tables to the text. The original paper contained two-tailed results even though the article's text described only an interest in one-tailed inferences. Here, we concern ourselves with the more appropriate one-tailed results in Table 1.

As seen in Table 1, Pearson's chi-square is more likely to indicate a difference than Fisher's test. The chi-square results seem to more intuitively follow the data structure, but they are not valid at the smaller sample sizes. Interestingly, Johns *et al.* concluded that there is a strong possibility that there exist greater respiratory hazards to mesquite broiler cooks, despite only detecting one significant difference using two-tailed Fisher's tests. Examination of the data is very useful for understanding whether an effect might exist, but a valid test is

TABLE 2

MORTALITY DATA FROM TWO ANAESTHESIA TECHNIQUES FOR ELDERLY PATIENTS RECEIVING EMERGENCY HIP SURGERY AND THE p VALUES FROM THREE ONE-TAILED TESTS

Anesthesia type	Result		Test results
	Alive	Death	
Spinal	34	3	Pearson chi-square $p = .037$
			Yates continuity $p = .070$
General	30	9	Fisher's "exact" $p = .069$

required to make more concrete inferences about the data. If these data were blindly analyzed with a canned program, one would be led to Fisher's test and, if two-tailed tests are performed, one might be led to believe that an effect did not exist.

We also present results from the unconditional test of McDonald *et al.* (8, 9). This test is one of a number of tests developed for analyzing 2×2 tables with small cell sizes (e.g., (2, 4)). It does not seem to be generally well known, although it is frequently referenced in statistical articles on analyzing 2×2 tables. The assumptions for this test are more easily met (in our small cell size situation) than for the Pearson chi-square and Fisher's "exact" test. However, this test is not incorporated into standard program packages and the user must rely on published tables to conduct the test. The results in Table 1 indicate that it is more likely to detect a difference than Fisher's test.

EXAMPLE 2

We now consider the data in Table 2. These data are from an article by Davis and Laurenson (10) where spinal anesthesia is compared to general anesthesia for elderly patients undergoing emergency hip surgery. The general hypothesis for the statistical tests would be that spinal anesthesia poses less risk to elderly patients than general anesthesia. This also implies a one-tailed test and, although not explicitly stated by the authors, this is what was performed based on reproducing their results.

We perform three tests on these data; Pearson chi-square, Yates continuity corrected chi-square, and Fisher's "exact." The one-tailed p -value results are also given in Table 2. The results in Table 2 exemplify the conservative nature of Fisher's exact test and Yates continuity corrected chi-square where very similar p values of .070 and .069, respectively, result. The p values for the Pearson chi-square are roughly only half as large. Davis and Laurenson (10) chose to use the Yates continuity corrected chi-square test even though the cell sizes are large enough (e.g., all expected values are greater than 5) that the common criteria for using Pearson's chi-square are satisfied.

In terms of reporting results the Fisher's "exact" test and Yates continuity

corrected chi-square produce results that could be considered "nearly" or "borderline" significant, whereas the Pearson chi-square (using the $p = .05$ criteria most frequently encountered in scientific journals) would be considered significant. The importance of the choice of test and reporting of the associated p value is further demonstrated by considering a survey paper contained in a book on anesthesia. Based on the results given in Davis and Laurensen (10) the survey paper by McLeskey (11) states simply that differences in mortality between the two anesthesia methods were not statistically significant. Had the original authors used the Pearson chi-square from which to base their inferences, the more general survey paper would probably have indicated that spinal anesthesia was a preferable treatment. This demonstrates that inferences are often interpreted and carried beyond the original report and this stresses the need for special attention to the use of the most appropriate analytical methods.

DISCUSSION

Several important considerations were presented here for analyzing data sets such as those in Tables 1 and 2. First, it is important to examine and understand the data rather than assume that a canned program would produce the appropriate analysis (we are not suggesting that this is the case for either example data set, but rather illustrating that this could happen). Second, the strengths and weaknesses of three of the most common methods for analyzing 2×2 tables are indicated. Third, it is not uncommon for data sets to fall in an area where well-known tests may not work well, and, if so, one must look for an alternative, which may not be part of a canned program. Fourth, the results reported may be carried well beyond the original outlet, which further stresses the need to apply the most appropriate procedure available.

The 2×2 contingency table is among the most common types of data set. They appear simple and much computer software is readily available for analysis, but selection of the appropriate analysis is a potential problem in many situations. It is the responsibility of the investigator to assure that the correct analysis and inferences are produced.

REFERENCES

1. YATES, F. Contingency tables involving small numbers and the χ^2 test. *J. R. Statist. Soc. Suppl.* **1**, 217 (1934).
2. D'AGOSTINO, R. B., CHASE, E., AND BELANGER, A. The appropriateness of some common procedures for testing the equality of two independent binomial proportions. *Am. Statist.* **42**, 198 (1988).
3. GRIZZLE, J. E. Continuity correction in the χ^2 test for 2×2 tables. *Am. Statist.* **21**, 28 (1967).
4. UPTON, G. J. G. A comparison of alternative tests for the 2×2 comparative trial. *J. R. Statist. Soc. Ser. A* **145**, 86 (1982).
5. DIXON, W. J., AND MASSEY, F. J., JR. "Introduction to Statistical Analysis," 3rd ed. McGraw-Hill, New York, 1969.
6. SNEDECOR, G. W., AND COCHRAN, W. G. "Statistical Methods," 7th ed. Iowa State Univ. Press, Ames, IA, 1980.

7. JOHNS, R. E., JR., LEE, J. S., AGANHIAN, B., GIBBONS, H. L., AND READING, J. L. Respiratory effects of mesquite broiling. *J. Occup. Med.* **28**, 1181 (1986).
8. McDONALD, L. L., AND MILLIKEN, G. A. "A Nonrandomized Test for Comparing Two Proportions." College of Commerce and Industry Research Paper 94, University of Wyoming, Laramie, WY, 1975.
9. McDONALD, L. L., DAVIS, B. M., AND MILLIKEN, G. A. A nonrandomized unconditional test for comparing two proportions in 2 × 2 contingency tables. *Technometrics* **19**, 145 (1977).
10. DAVIS, F. M., AND LAURENSEN, V. G. Spinal anaesthesia or general anesthesia for emergency hip surgery in elderly patients. *Anaesthesia Intensive Care*, **9**, 52 (1981).
11. McLESKEY, C. H. Anesthesia for the geriatric patient. In "Advances in Anesthesia" (Stoelting, Ed.), pp. 31-68. Year Bk Med., Chicago, 1985.